外国手話データセットを活用した 日本手話動画からの音節構成要素認識

木全 純大　三輪 誠　佐々木 裕　原 大介
豊田工業大学
{sd20038,makoto-miwa,yutaka.sasaki,daisuke}@toyota-ti.ac.jp

概要

本研究では日本手話の「手の位置」・「手の動き」・「手型」から構成される音節構成要素を深層学習を用いて自動認識するシステムの構築において，限られた日本手話動画データから音節構成要素を高い性能で認識するため，日本手話と外国手話の言語間の共通の特徴を活かす転移学習手法を提案する．「手の位置」・「手の動き」・「手型」の認識では，マルチタスク学習を行なった．利き手における音節構成要素を対象とした実験では，原らが作成した日本手話データセットにおいて，本稿で提案する外国手話データセットを利用した転移学習が音節構成要素認識の性能向上に繋がることを明らかにした．

1　はじめに

日本手話は日本語と異なる自然言語であり，手指動作と非手指動作から構成される．手指動作は「手の位置」・「手の動き」・「手型」から構成され，これらは，日本手話の音節に相当する単位を構成する要素（**音節構成要素**）である．一方で，非手指動作には顔の表情や視線などの文法要素が含まれる．原は，音節の構成規則を解析するため，対応する音節の動画を撮影・収集し，音節を音節構成要素に分解・記号化したデータベースを作成した [1]．現在，日本手話動画から自動で十分な性能で音節構成要素を認識するシステムはない．このような音節構成要素の自動での認識ができれば，まだ分析されていない多くの音節に対応して，データベースを人手で拡張するコストを減らすための補助や自動での拡張が可能となる．

近年，手話研究において，手話動画から gloss[1)] を予測する手話認識 [2, 3] など，動画から時空間情報を抽出可能な深層学習による動画解析手法を利用した手法が提案されている．深層学習で高い性能を実現するには，ラベル付けされた大量の訓練データが必要となるが，音節構成要素がラベル付された日本手話動画データの数は限られている．一方で，アメリカ手話などの外国手話のデータは多く存在し，外国手話と日本手話には，手指動作と非手指動作により手話を表現する共通の特徴が存在する．転移学習 [4] などにより，この言語間の共通点を活かすことで，データ数の限られた日本手話における認識性能の向上につながることが期待できるが，我々の知る限りこのような研究は行われていない．

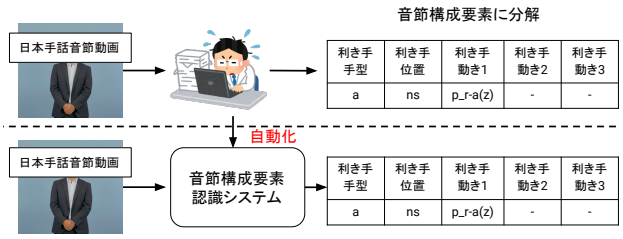

図 1: 提案する日本手話動画から音節構成要素を自動で認識するシステム

そこで，本研究では，日本手話動画からの音節構成要素認識に向けて，外国手話データセットを活用した転移学習の有効性を検証する．本稿では，簡単のため，利き手における「手の位置」・「手の動き」・「手型」を対象に評価・検証を行う．具体的には，外国手話データセットを利用して事前学習した深層学習モデルのパラメタを初期値として，モデルを日本手話動画データで学習し，音節構成要素の認識を行う．また，「手の位置」・「手の動き」・「手型」の認識において，分類のベースとなるモデルを共有し，同時に学習するマルチタスク学習を導入し，その有効性を調査する．本研究の貢献は次の通りである．

- 日本手話動画から自動で音節構成要素を認識するシステムを構築した．
- データの限られた日本手話において，転移学習

1） gloss は，手話の意味に対応する音声言語の単語を示す．

により外国手話データで学習したモデルを利用することの有効性を初めて示した．
- 日本手話の音節構成要素のマルチタスク学習を通じて，各タスク間の情報共有が必ずしも認識性能の向上に寄与しないことを示した．

2 関連研究

2.1 日本手話データセット

長嶋らは言語学や工学などの様々な分野の研究者が共通に利用できる多用途型日本手話データベースを構築した [5]．様々な観点からの分析に対応するため，データベースには，高解像度のカメラで日本手話者母語話者の動作を撮影した映像データに加えて，3 次元動作データ・深度データ，さらにこれらのデータを同期収録したデータが含まれている．現在，4,873 個の gloss と 10 個の対話を収録したデータが提供されている．

原は，日本手話母語話者の協力を得て，日本手話音節を表現した動画を撮影・収集し，対応する音節を日本手話コーディングマニュアル [1] に従い，「手の位置」・「手の動き」・「手型」からなる音節構成要素に分解・記号化したデータベースを作成した．撮影した音節動画は合計 1,086 本あり，各動画には約 300 フレーム含まれている．「手の位置」は，手が空間または身体のどの位置にあるかを示し，22 種類に分類される．「手の動き」は，手の移動の仕方を 55 種類に区別し，「手型」は 69 種類に分けられる．データベースには，利き手と非利き手の「手の位置」・「手の動き」・「手型」に加えてそれぞれの要素を詳細に記載しており，合計で 27 種類の要素が含まれている．

2.2 機械学習・深層学習を利用した手話研究

コンピュータビジョン分野において，gloss を予測する手話認識や手話を音声言語に翻訳する手話翻訳などの機械学習・深層学習を用いた手話研究 [2, 6] が盛んに行われている．Jiang らは，動画やキーポイント[2)]情報に加えて身体情報・動作情報・深度情報を統合したフレームワークを提案した [3, 7]．また，Zuo らは，64 フレームと 32 フレームの動画とキーポイントを用いて異なる時間情報を考慮した Video-Keypoint Network (VKNet) を提案した [2]．VKNet は VKNet-64 と VKNet-32 の 2 つのサブネットワークから構成されており，さらに動画とキーポイントのエンコーダがそれぞれのサブネットワークに含まれる．エンコーダには，時空間情報を考慮できる 3D Convlutional Neural Network の一つである S3D [8] を用いている．動画から学習済みの姿勢推定モデル [9] を使用してキーポイントを推定した後，64 フレームと 32 フレームの動画とキーポイントをそれぞれ VKNet-64 と VKNet-32 に入力する．各ネットワークから得られた表現ベクトルを結合した表現ベクトルは，gloss の予測に利用される．VKNet は，手話認識のための複数のデータセットにおいて優れた性能を発揮した．

音節構成要素を考慮した手話研究も行われている．[10, 11, 12, 13, 14]．アメリカ手話において，Kezar らは，gloss に加えて 16 種類の音節構成要素を手話動画にラベル付けした大規模なデータセットを構築し，音節構成要素認識を通してその特徴を学習することが手話認識の性能向上に繋がると示した [12]．また，日本手話においては，波多野らが，機械学習手法を利用して「手の位置」・「手の動き」・「手型」の認識を行い，各認識スコアの重み付け和に基づいて gloss を認識する手話認識システムを構築した [13]．この手法は認識のための特徴を手動で作成する必要がある．

3 提案手法

本研究では，日本手話動画からの音節構成要素認識に向けて，外国手話データセットを利用した転移学習手法を提案する．簡単のため，2.1 節で説明した 27 種類の要素のうち，利き手の「手の位置」・「手の動き」・「手型」を対象とする．提案モデルの全体像を図 2 に示す．外国手話データセットで事前学習した 2.2 節で説明した VKNet のパラメタを初期値として，マルチタスク学習による日本手話動画からの利き手の音節構成要素認識を行う．

日本手話動画からの利き手の音節構成要素認識では，「手の位置」・「手の動き」・「手型」をそれぞれ分類するマルチタスク学習に取り組む．2.1 節で説明したように，「手の位置」・「手の動き」・「手型」はそれぞれ 22，55，69 種類ある．各音節構成要素の分類を実現するために，外国手話データセットで事前学習した VKNet に，「手の位置」・「手の動き」・「手型」に対応する 3 つの全結合層を新たに追加した音節構成要素認識モデルを作成する．「手の位置」・「手型」

2) キーポイントは，画像や動画中の，目・手・関節などの人体部位の座標を示す．

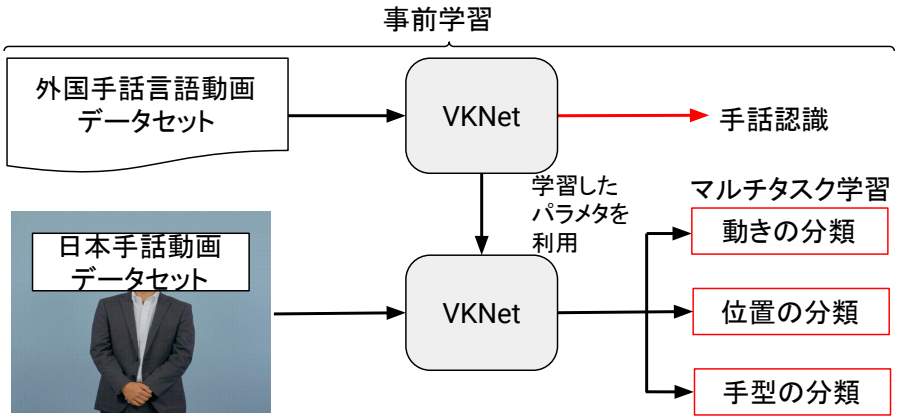

図 2: 提案手法の概要

は日本手話動画に複数の種類の中から一つのラベルが付与されているので，全結合層の出力ベクトルに softmax 関数を適用し，予測確率が一番高いクラスをモデルの予測とする多クラス分類を行う．「手の動き」に関しては，日本手話動画に複数の動きの種類がラベル付けされているので，全結合層の出力ベクトルに sigmoid 関数を適用することでそれぞれの動きの種類ごとに 2 クラス分類を行い，予測確率が閾値よりも大きい動きの種類をモデルの予測とする多ラベル分類を行う．

学習における損失については，「手の位置」・「手型」の分類損失 $L_{position}$, L_{shape} にはクロスエントロピー損失を用いる．また，「手の動き」の分類損失 $L_{movement}$ には Ridnik [15] らが提案した asymmetric loss 関数を用いる．「手の動き」に関しては，55 種類それぞれについて 2 クラス分類を行うが，1 本の動画で同時に行われる動きは最大で 3 つまでである．そのため，正例が少なく，負例が多い偏った分類問題となる．この損失を用いることで，多ラベル分類においてラベルの偏りに対処し，適切に正例を学習できるようにする．学習においては，タスク間の情報共有による各音節構成要素の認識性能の向上を期待して，マルチタスク学習を行う．具体的には，VKNet を共有し，各音節構成要素の分類損失の和を学習時の損失関数 L とする．

$$L = L_{position} + L_{movement} + L_{shape} \tag{1}$$

4 実験設定

2.1 節で説明した原が作成したデータベースを用いて，提案手法の評価を行った．学習・評価のために，欠損値を除いた 1,072 本の音節動画データをランダムに分割して，750・161・161 本のデータをそれぞれ訓練・開発・テストデータとして使用した．データに含まれる利き手の「手の位置」・「手の動き」・「手型」の分布を付録 A に示す．付録 A に記載されているように日本手話データセットは不均衡データセットである．極端に難しい分類問題を排除するため，学習データに含まれる各音節構成要素のクラスに対応するデータ数が 5 未満の場合，学習に含めず，偽陰性の予測として扱うこととした．評価指標としてはマイクロ F 値を採用した．

事前学習で使用した手話認識のデータセットはアメリカ手話の Word-Level American Sign Language (WLASL) [16] データセットを用いた．WLASL には，2,000 個の gloss と 14,289・3,916・2,878 本の訓練・開発・テストデータからなる計 21,083 本の動画データが含まれている．WLASL データセットで学習済みの VKNet[3] を事前学習後のパラメタとして利用した．

日本手話の音節構成要素認識においては，外国手話データセットを利用した転移学習の有効性を調べるため，WLASL データセットで学習済みのモデルのパラメタを初期値とする VKNet と，ランダムに設定したパラメタを初期値とする VKNet で音節構成要素認識の実験を行い，認識性能を比較した．また，マルチタスク学習においては，タスク間の情報共有が認識性能に与える影響を明らかにする目的で，各タスクに同時に取り組む場合とそれぞれのタスクを独立に取り組む場合の認識性能を比較した．

最適化手法として Adam[17] を用い，学習率は 5×10^{-5} に設定した．また，学習率をエポックごとに変化させるスケジューラとしてコサインアニーリ

3) https://github.com/FangyunWei/SLRT/tree/main/NLA-SLR

表 1: 転移学習の有無・マルチタスク学習の有無における音節構成要素認識の結果
（評価指標はマイクロ F 値 [%]．3 回の評価の平均と標準偏差を示した．）

Method		開発データ			テストデータ		
		「手の位置」	「手の動き」	「手型」	「手の位置」	「手の動き」	「手型」
Multitask	VKNet	82.40 ± 1.27	34.06 ± 0.52	39.75 ± 1.83	80.33 ± 1.06	38.55 ± 1.25	35.20 ± 2.55
	+ pretrained	82.20 ± 1.17	39.94 ± 2.85	47.41 ± 2.29	81.99 ± 0.00	45.76 ± 0.82	42.23 ± 1.34
Singletask	VKNet	83.85 ± 1.01	34.57 ± 0.39	43.89 ± 0.29	80.75 ± 1.02	38.29 ± 2.54	39.54 ± 1.05
	+ pretrained	83.44 ± 0.77	44.98 ± 1.06	47.82 ± 1.02	81.16 ± 2.05	52.41 ± 0.86	44.72 ± 3.55

ングを使用した．過学習を抑制するため，ドロップアウト [18] と正則化を適用し，それぞれの値を 0.2 と 10^{-3} に設定した．実験環境の詳細を付録 B に示した．

5 結果・考察

5.1 結果

外国手話データセットを活用した転移学習の有無・マルチタスク学習の有無による日本手話動画からの音節構成要素認識の結果を表 1 に示す．開発データとテストデータに含まれている「手の位置」・「手の動き」・「手型」の分布が異なることを考慮して，両データでの評価を載せた．WLASL データセットで事前学習したパラメタを初期値とする VKNet を用いて音節構成要素認識を行った結果，ランダムに設定したパラメタを初期値とする VKNet を用いたときと比較して，マルチタスク学習の有無に関係なく，「手の動き」と「手型」の認識性能が大きく向上した．「手の位置」の認識性能はほとんど変わらなかったが，提案した外国手話データセットを活用した転移学習が日本手話の音節構成要素認識に有効であることが確認できた．一方で，マルチタスク学習を実施したときには，「手の位置」・「手の動き」・「手型」を独立に学習した場合と比較して，音節構成要素の認識性能がほとんど変化しない，または低下するという結果になった．

5.2 考察

転移学習が VKNet の音節構成要素予測に与える影響を考察するために，説明可能な AI 技術である Adaptive Occlusion Sensitivity Analysis (AOSA) [19] を用いて，VKNet が動画内のどの部分に着目して音節構成要素の予測を行ったのか，その予測根拠を可視化した．付録 C の図 3 および図 4 では，同じ動画で転移学習を行った場合に「手の位置」・「手の動き」・「手型」の全て正解した事例と、転移学習を行わない場合に，「手の動き」が不正解だった事例において、AOSA によって得られた VKNet の予測根拠の違いを示す．転移学習の適用により，VKNet は手話者の利き手である右手に注目して音節構成要素の認識を行うようになっており，転移学習を行わない場合に比べて正確な予測ができるようになったと考えられる．

6 おわりに

本研究では，日本手話動画からの音節構成要素認識における外国手話データセットを活用した転移学習の有効性の検証を目的に，外国手話データセットで事前学習したモデルのパラメタを初期値として日本手話動画データで学習する音節構成要素の認識手法を提案した．また，音節構成要素認識におけるマルチタスク学習の影響についても評価を行った．日本手話データセットにおける評価から，外国手話データセットを利用した転移学習が限られた日本手話動画データからの音節構成要素認識に有効であること，マルチタスク学習によるタスク間の情報共有が必ずしも認識性能改善に有効でないことが分かった．また，VKNet の予測根拠を説明可能な AI 技術である AOSA を用いて可視化することにより，転移学習が音節構成要素予測に与える影響を調査した．

今後の課題として，一つのモデルで複数のタスクを高い認識性能で解くことができるように，カリキュラム学習などの学習方法を検討する．また，利き手だけでなく非利き手の音節構成要素の認識にも取り組む．

謝辞

本研究は JSPS 科研費　JP23H00626 の助成を受けたものです.

参考文献

[1] 原大介. 新日本手話コーディングマニュアル. 2019.

[2] Ronglai Zuo, Fangyun Wei, and Brian Mak. Natural language-assisted sign language recognition. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 14890–14900, June 2023.

[3] Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. Skeleton aware multi-modal sign language recognition. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops**, 2021.

[4] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. **Proceedings of the IEEE**, Vol. 109, No. 1, pp. 43–76, 2021.

[5] 長嶋, 原, 堀内, 酒向, 渡辺, 菊澤, 加藤, 市川. 多様な研究分野に利用可能な超高精細・高精度手話言語データベースの開発. 第 3 巻, pp. 148–155. 国立国語研究所, 2018.

[6] Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. A simple multi-modality transfer learning baseline for sign language translation. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 5120–5130, 2022.

[7] Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. Sign language recognition via skeleton-aware multi-model ensemble. **arXiv preprint arXiv:2110.06161**, 2021.

[8] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In **Proceedings of the European conference on computer vision (ECCV)**, pp. 305–321, 2018.

[9] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, June 2019.

[10] Federico Tavella, Viktor Schlegel, Marta Romeo, Aphrodite Galata, and Angelo Cangelosi. WLASL-LEX: a dataset for recognising phonological properties in American Sign Language. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 453–463, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[11] Kezar. Improving sign recognition with phonology. In **Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics**, pp. 2732–2737. Association for Computational Linguistics, May 2023.

[12] Lee Kezar, Elana Pontecorvo, Adele Daniels, Connor Baer, Ruth Ferster, Lauren Berger, Jesse Thomason, Zed Sevcikova Sehyr, and Naomi Caselli. The sem-lex benchmark: Modeling asl signs and their phonemes. 2023.

[13] 美歌波多野, 慎司酒向, 正北村. Kinect v2 による手話動作の 3 要素に基づく実時間手話認識. 電子情報通信学会技術研究報告 = IEICE technical report : 信学技報, Vol. 115, No. 491, pp. 59–64, 03 2016.

[14] 光希有賀, 慎司酒向, 正北村. 日本手話の音韻構造を考慮した hmm に基づく手話認識. 電子情報通信学会技術研究報告 = IEICE technical report : 信学技報, Vol. 110, No. 220, pp. 127–132, 10 2010.

[15] Emanuel Ben-Baruch, Tal Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification, 2020.

[16] DONGXU LI, Cristian Rodriguez, Xin Yu, and HONGDONG LI. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In **Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)**, March 2020.

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, **3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings**, 2015.

[18] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. **Journal of Machine Learning Research**, Vol. 15, No. 56, pp. 1929–1958, 2014.

[19] Tomoki Uchiyama, Naoya Sogi, Koichiro Niinuma, and Kazuhiro Fukui. Visually explaining 3d-cnn predictions for video classification with an adaptive occlusion sensitivity analysis. In **Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision**, pp. 1513–1522, 2023.

[20] FirstName LastName. Title of the article. **Journal of Natural Language Processing**, Vol. 13, No. 1, pp. 251–258, 2006.

A　日本手話データセットの統計

日本手話データセットに含まれる「手の位置」・「手の動き」・「手型」の分布を表 2 に示す.

B　実験環境

本実験を行った環境について説明する. Python [20] のバージョン 3.10 と深層学習ライブラリである PyTorch [20] のバージョン 2.0.0 を用いて実装した.

C　AOSA を用いた VKNet の予測根拠の可視化

VKNet にとって，動画（画像）中で音節構成要素の認識に重要な箇所は赤く，それ以外の領域は青く表示される．転移学習を行わないときと行なったときの VKNet の予測根拠を可視化した図をそれぞれ図 3 と図 4 に示した.

図 3: 転移学習を行わないときの VKNet の予測根拠を可視化

図 4: 転移学習を行なったときの VKNet の予測根拠を可視化

表 2: 「手の位置」・「手の動き」・「手型」のデータ数

手の動き	データ数	手型	データ数	手の位置	データ数
p_r	142	1	138	ns	835
p_o	135	b	125	uf(kmk)	40
or_s	120	s	57	lf(k)	32
p_d	117	f	55	tk(u)	23
hs_f	80	5	53	uf(dk)	22
hs_e	77	a	48	uf(m)	17
p_c(xy)	69	L-f	42	fc	16
p_u	64	b-f	40	el	13
p_l	61	u	40	ns(u)	13
p_7	51	o	34	tk	12
p_t	47	L-b	33	mf	9
or_p	47	v	31	lf(g)	8
p_c(xz)	39	5-b	30	lf(h)	8
or_e	28	L	27	la	5
hs_e/f	26	b(lax)	26	ua	5
p_o=d	26	w	26	nk	4
hs_wiggle	23	y	25	am	3
p_c(yz)	23	1-b	19	er	3
p_x	22	b4	19	lg	1
p_r=u	21	ko	17	uf(k)	1
p_u/d	20	u-b	17	ow	1
dot	19	b-f(q)	16	fc → tk	1
p_l=d	19	1(nf)	12		
or_s/p	18	7	12		
or_trill	16	c	11		
or_f	16	i	11		
p_r=o	15	L-f(lax)	10		
p_s/s	13	L-f(q)	9		
p_o/t	12	v-b	9		
p_r=d	11	5(lax)	7		
p_o=u	10	b-b	7		
p_l=u	8	k	7		
or_e/f	8	t	7		
hs_w	8	78-f	5		
hs_trill	7	8-f	5		
p_t=u	6	r	4		
p_trill	5	8(s)-b	3		
or_c	5	4	2		
p_l=o	5	b(lax) → 5	2		
or_sb/pb	4	ch	2		
p_r=t	4	1 → 4	1		
hs_f(c)	3	1 → L-f	1		
p_r=u/l=d	2	1 → a	1		
nm	2	1 → i	1		
hs_e(c)	2	2	1		
p_u=t	2	3	1		
trill_p	1	7-b	1		
hs_1 → i	1	7-f	1		
hs_1 → a	1	78-b	1		
p_???	1	78-f(q)	1		
hs_L	1	8(s)	1		
hs_n	1	L-b → b	1		
hs_L → 7	1	L-f(q) → L-f(q)	1		
p_t=d	1	L-f → a	1		
hs	1	L → 7	1		
		a(nf)	1		
		a-b	1		
		b-f(q) → b-f	1		
		b-f(q) → s	1		
		c → o	1		
		horn	1		
		i-b	1		
		o → 5-b	1		
		s → w	1		
		ts	1		
		u-f	1		
		w(nf)	1		
		w-b	1		
		薬	1		